

ACHIEVING PRIVACY PRESERVING CLUSTERING IN IMAGES THROUGH MULTIDIMENSIONAL SCALING

T. SUDHA¹ & P. NAGENDRA KUMAR²

¹Professor, Department of Computer Science, Sri Padmavathi Mahila University, Tirupati,
Andhra Pradesh, India

²Research Scholar, Department of Computer Science, Vikrama Simhapuri University, SPSR
Nellore, Andhra Pradesh, India

ABSTRACT

Privacy preserving data mining has become one of the recent trends in research. This paper makes a comparative study of the dimensionality reduction techniques such as Singular value decomposition, Principal component analysis and Multi dimensional scaling in the context of privacy preserving clustering. High-dimensional data such as images have been reduced to lower dimensional data through techniques such as Singular value decomposition, Principal component analysis and Multidimensional scaling. Cluster analysis has been performed on the original data and as well as on all of the three lower dimensional data obtained through Singular value decomposition, Principal component analysis and Multidimensional scaling using K-Means algorithm with varying number of clusters. Mean squared error has been considered as one of the parameters for comparison. The experimental results show that the mean squared error obtained on the original data is almost same as the mean squared error obtained on the data reduced through Multidimensional scaling, but the order of the values differs due to the random selection of cluster centers. Space has also been considered for comparison. The results show that the space required by the lower dimensional data obtained through multidimensional scaling is far less than the space required by the original data. It is also clearly evident that privacy preserving clustering of pixels can be achieved through multidimensional scaling with less amount of storage space.

KEYWORDS: Privacy Preserving Clustering, Singular Value Decomposition, Principal Component Analysis, Multidimensional Scaling

Received: Dec 11, 2015; **Accepted:** Dec 18, 2015; **Published:** Dec 29, 2015; **Paper Id.:** IJCSEITRFEB20162

INTRODUCTION

Data mining refers to extraction of non-trivial, implicit, previously unknown and potentially useful information or patterns from data in large databases. Data mining, with its promise to efficiently discover valuable, non-obvious information from large databases, is particularly vulnerable to misuse. Hence data mining with privacy concerns known as privacy preserving data mining has been developed. Privacy preserving data mining refers to safeguard sensitive information from unsolicited or unsanctioned disclosure. Privacy preserving data mining has numerous applications such as Medical databases, home land security applications, genomic privacy and bioterrorism applications. Privacy preserving clustering is one of the functionalities of privacy preserving data mining. The goal of privacy preserving clustering is to protect the underlying attribute values of objects subjected to cluster analysis.

Image processing is any form of signal processing for which the input is an image and the output may be

an image or a set of characteristics or parameters related to the image. Privacy preserving is of primary concern in images in order to protect the information from the images. Images are normally considered as high dimensional data as they require large amount of storage space. Dimensionality reduction techniques can be used to reduce the high dimensional data to low dimensional data. Dimensionality reduction can be defined as the search for a low-dimensional manifold that embeds the high dimensional data. Many dimensionality reduction techniques such as Singular value decomposition, Principal component analysis, Multidimensional scaling, Independent component analysis, Projection pursuit etc have been developed. The present work focuses on three dimensionality reduction techniques such as Singular value decomposition, Principal Component analysis and Multidimensional scaling.

Singular Value Decomposition (SVD)

It can be seen as a method for data reduction because if we have identified where the most variation is, it is possible to find the best approximation of the original data points using fewer dimensions. SVD is based on a theorem from linear algebra which says that a rectangular matrix A can be broken down into the product of three matrices.

- An orthogonal matrix U
- A diagonal matrix S
- The transpose of the orthogonal matrix V

The theorem is usually represented as

$$A_{mn} = U_{mn} S_{nn} V_{nn}^T$$

$$\text{Where } UU^T = I \quad V^T V = I$$

The columns of U are orthonormal Eigen vectors of AA^T

The columns of V are orthonormal Eigen vectors of $A^T A$

S is a diagonal matrix containing the square roots of Eigen values from U or V in descending order.

Principal Component Analysis

It is a way of identifying patterns in data and expressing the data in such a way as to highlight their similarities and differences.

- Get some data.
- Subtract the mean.
- Calculate covariance matrix.
- Calculate the Eigen vectors and Eigen values of the covariance matrix.
- Choose components and form a feature vector. Feature vector = $(\text{eigen}_1, \text{eigen}_2, \dots, \text{eigen}_n)$
- Derive the new data set.

Multidimensional Scaling

Multidimensional scaling (MDS) is a set of related statistical techniques often used in information visualization

for exploring similarities or dissimilarities in data. MDS is a special case of ordination. An MDS algorithm starts with a matrix of item–item similarities, and then assigns a location to each item in N -dimensional space, where N is specified a priori. For sufficiently small N , the resulting locations may be displayed in a graph or 3D visualization.

The data to be analyzed is a collection of I objects (colors, faces, stocks, . . .) on which a distance function is defined,

$\delta_{i,j} :=$ distance between i^{th} and j^{th} objects.

These distances are the entries of the dissimilarity matrix

$$\Delta := \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,I} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,I} \\ \vdots & & & \\ \delta_{I,1} & \delta_{I,2} & \cdots & \delta_{I,I} \end{pmatrix}$$

The goal of MDS is, given Δ , to find I vectors $x_1, \dots, x_I \in \mathbf{R}^N$ such that

$$\|x_i - x_j\| \approx \delta_{i,j} \text{ for all } i, j \in I,$$

Where $\|\cdot\|$ is a vector norm. In classical MDS, this norm is the Euclidean distance, but, in a broader sense, it may be a metric or arbitrary distance function.

In other words, MDS attempts to find an embedding from the I objects into \mathbf{R}^N such that distances are preserved. If the dimension N is chosen to be 2 or 3, we may plot the vectors x_i to obtain a visualization of the similarities between the I objects. Note that the vectors x_i are not unique: With the Euclidean distance, they may be arbitrarily translated, rotated, and reflected, since these transformations do not change the pair wise distances $\|x_i - x_j\|$.

Present Work

Example 1

An image of 177×284 is considered and then the intensity of each pixel value is retrieved and stored in a matrix. It is treated as a higher dimensional data. Then this matrix has been reduced to lower dimensional data using three dimensionality reduction techniques such as singular value decomposition, principal component analysis and multi dimensional scaling. Clustering is performed on the original data and as well as the three lower dimensional data obtained through dimensionality reduction techniques using K-means algorithm of MATLAB with varying number of clusters. Mean squared error obtained from all the four datasets (original data set and three reduced data sets) is tabulated. From the tabulated values it is very clearly evident that clustering performed on the data obtained through Multidimensional scaling is same as the clustering performed on the original data but the order of the mean squared error values differ due to the random selection of cluster centers.

Table 1: Analysis of Mean Squared Error

Number of Clusters	Mean Squared Error Obtained From Original Data	Mean Squared Error Obtained After Applying SVD On the Original Data	Mean Squared Error Obtained After Applying PCA on the Original Data	Mean Squared Error Obtained After Applying MDS on the Original Data
K=2	1.0e+003*1.3796 1.0e+003*0.7125	0 1.0e+003*2.0924	1.0e+003*0.4697 1.0e+003*1.6163	1.0e+003*0.5508 1.0e+003*1.5742
K=3	926.1486 296.3791 449.6284	0 0 1.0e+003*1.6090	867.7118 492.0208 373.8865	469.9517 449.6284 774.2366
K=4	387.8102 359.9119 434.4420 211.4105	0 0 0 1.0e+003*1.2393	359.9119 211.4105 387.8102 434.4420	434.4420 387.8102 359.9119 211.4105
K=5	248.7373 197.9311 187.9306 219.4571 388.1285	0 1.0e+003*1.0197 0 0 0	96.4421 269.7617 149.9284 434.4420 347.2754	187.9306 219.4571 387.8102 235.4974 211.4105
K=6	203.3002 223.5007 186.6821 187.9306 139.0233 219.4571	0 0 0 0 852.3622 0	223.5007 187.9306 219.4571 186.6821 139.0233 203.3002	47.0881 87.5181 267.9239 133.5527 604.1930 105.7266
K=7	325.2469 199.9317 61.3569 80.0228 203.4134 124.7103 47.0881	759.4627 0 0 0 0 0 0	46.6407 44.7268 135.3357 434.4420 347.2754 38.5668 105.7266	63.0925 187.9306 124.7103 219.4571 139.8104 47.0881 300.9813

Table 2: Analysis of Space

Space Required by Original Matrix	Space Required by Reduced Matrix Obtained Through SVD	Space Required by Reduced Matrix Obtained Through PCA	Space Required by Reduced Matrix Through MDS
402144 bytes	3264 bytes	374964 bytes	249216 bytes

Analysis of Privacy Preserving Nature

The original image and then the images obtained through SVD, PCA and MDS are analyzed. It is clearly evident that we can obtain privacy preserving clustering in images through Multidimensional scaling.

Figure 4: After Applying MDS, the Image Looks Like**Example 2**

An image of 1600×1024 is considered and then the intensity of each pixel value is retrieved and stored in a matrix. It is treated as a higher dimensional data. Then this matrix has been reduced to lower dimensional data using three dimensionality reduction techniques such as singular value decomposition, principal component analysis and multi dimensional scaling. Clustering is performed on the original data and as well as the three lower dimensional data obtained through dimensionality reduction techniques using K-means algorithm of MATLAB with varying number of clusters. Mean squared error obtained from all the four datasets (original data set and three reduced data sets) is tabulated. From the tabulated values it is very clearly evident that clustering performed on the data obtained through Multidimensional scaling is same as the clustering performed on the original data but the order of the mean squared error values differ due to the random selection of cluster centers.

Table 3: Analysis of Mean Squared Error

Number of Clusters	Mean Squared Error Obtained From Original Data	Mean Squared Error Obtained After Applying SVD on the Original Data	Mean Squared Error Obtained After Applying PCA on the Original Data	Mean Squared Error Obtained After Applying MDS on the Original Data
K=2	1.0e+004*1.2620	1.0e+004*0	1.0e+004*2.3668	1.0e+004*2.3668
	1.0e+004*2.3668	1.0e+004*3.7070	1.0e+004*1.2620	1.0e+004*1.2620
K=3	1.0e+004*1.2691	1.0e+004*0	1.0e+004*1.1351	1.0e+004*0.9859
	1.0e+004*1.0068	1.0e+004*0	1.0e+004*0.6907	1.0e+004*1.2691
	1.0e+004*0.8239	1.0e+004*2.8541	1.0e+004*1.2966	1.0e+004*0.8449
K=4	1.0e+003*6.5107	1.0e+004*2.3609	1.0e+003*7.0297	1.0e+004*2.2956
	1.0e+003*6.8864	1.0e+004*0	1.0e+003*6.8864	1.0e+004*0.8026
	1.0e+003*6.6082	1.0e+004*0	1.0e+003*6.6082	1.0e+004*0.1257
	1.0e+003*7.0297	1.0e+004*0	1.0e+003*6.5107	1.0e+004*0.1680
K=5	1.0e+003*5.0692	1.0e+004*0	1.0e+004*0.3319	1.0e+003*5.0692
	1.0e+003*6.6082	1.0e+004*0	1.0e+004*0.5672	1.0e+003*6.6082
	1.0e+003*2.3322	1.0e+004*2.0401	1.0e+004*1.1765	1.0e+003*2.3322
	1.0e+003*5.8904	1.0e+004*0	1.0e+004*0.0124	1.0e+003*5.9460
	1.0e+003*5.9460	1.0e+004*0	1.0e+004*0.6608	1.0e+003*5.8904
K=6	1.0e+003*4.1468	1.0e+004*0	1.0e+003*0.8774	1.0e+003*5.0291
	1.0e+003*2.3723	1.0e+004*0	1.0e+003*6.4952	1.0e+003*2.2275
	1.0e+003*1.0172	1.0e+004*1.8823	1.0e+003*6.4019	1.0e+003*6.6686
	1.0e+003*4.3285	1.0e+004*0	1.0e+003*2.4575	1.0e+003*4.5617
	1.0e+003*9.3882	1.0e+004*0	1.0e+003*3.2224	1.0e+003*1.0172
	1.0e+003*4.9953	1.0e+004*0	1.0e+003*4.5227	1.0e+003*4.8329
Table 3: Contd.,				
K=7	1.0e+003*0.1499	1.0e+004*0	1.0e+003*6.6082	1.0e+003*1.6636
	1.0e+003*0.6878	1.0e+004*0	1.0e+003*2.8006	1.0e+003*2.2275
	1.0e+003*4.5617	1.0e+004*0	1.0e+003*1.6403	1.0e+003*3.1532
	1.0e+003*1.5334	1.0e+004*0	1.0e+003*4.0652	1.0e+003*4.5022
	1.0e+003*5.3123	1.0e+004*0	1.0e+003*4.5022	1.0e+003*1.9198
	1.0e+003*6.6559	1.0e+004*0	1.0e+003*1.4731	1.0e+003*2.7218
	1.0e+003*5.4569	1.0e+004*1.7696	1.0e+003*1.8710	1.0e+003*6.6082

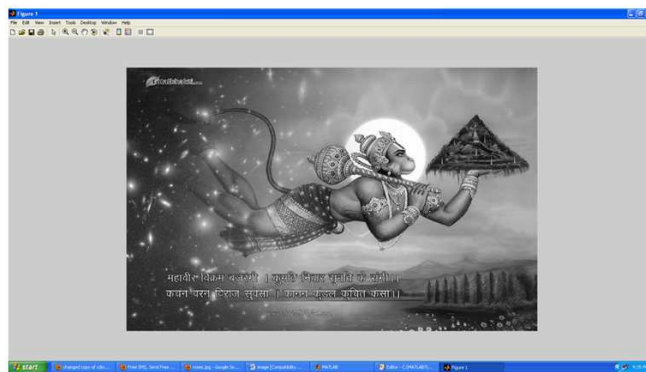
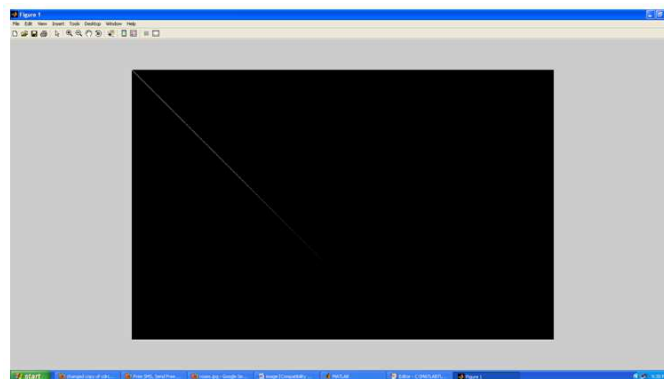
Table 4: Analysis of Space

Space Required by Original Matrix	Space Required by Reduced Matrix Obtained through SVD	Space Required by Reduced Matrix Obtained through PCA	Space Required by Reduced Matrix through MDS
13107200 bytes	18692 bytes	12577028 bytes	8380416 bytes

Analysis of Privacy Preserving Nature

The original image and then the images obtained after applying SVD, PCS and MDS are displayed below.

From the images displayed below, it is clearly evident that we can obtain privacy preserving clustering in images through Multi dimensional scaling.

**Figure 5: Original Image****Figure 6: After Applying SVD on the Original Image, the Image Looks Like****Figure 7: After Applying PCA on the Original Image, the Image Looks Like**

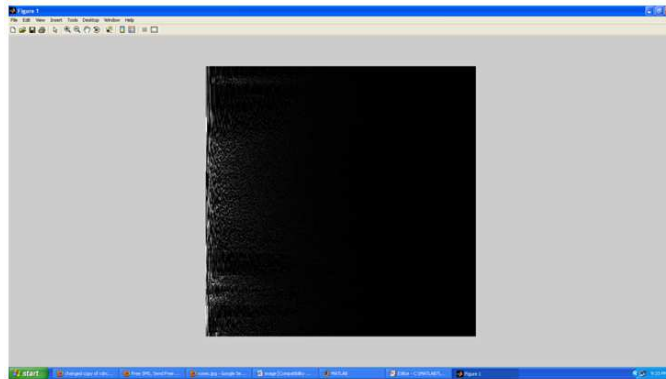


Figure 8: After Applying MDS on the Original Image, the Image Looks Like

CONCLUSIONS

It is very difficult and time consuming to perform clustering on a higher dimensional data due to large number of dimensions. If the higher dimensional data can be reduced to lower dimensional data, clustering can be performed easily. In order to convert the higher dimensional data in to lower dimensional data, a number of dimensionality reduction techniques have been developed. In this work we have been interested in three dimensionality reduction techniques and they are singular value decomposition, principal component analysis and multidimensional scaling. Different types of higher dimensional data have been considered and it has been reduced to lower dimensional data through the three above mentioned dimensionality reduction techniques. Cluster analysis has been done on the original data as well as the three lower dimensional data obtained through reduction techniques. It is very clear from the tabulated values that cluster analysis performed on the original data is almost same as the cluster analysis performed on the data obtained through multidimensional scaling and hence we can achieve privacy preserving clustering. Through Multidimensional scaling, we can achieve privacy preserving clustering with less amount of storage space. This work can be extended to other dimensionality reduction techniques.

REFERENCES

1. Ronald L Breiger, Scott A Boorman, Phipps Arabie (August 1975) : "An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling". *Journal of Mathematical Psychology*, Volume 12, Issue 3, Pages 328- 333
2. Jan de Leeuw, Patrick Mair (August 2009) : "Multidimensional scaling using majorization: SMACOF in R". *Journal of Statistical Software*, Issue 3, Volume 31
3. Michael D.Lee (2001) : "Determining the dimensionality of multidimensional scaling representations for cognitive modeling". *Journal of Mathematical Psychology*, Volume 45, 149-166
4. Robert B.Schneider (1992) : "A uniform approach to multidimensional scaling". *Journal of Classification*, Volume 9, Issue 2, pages 257-273
5. Rubner. Y, Tomasi.C, Guibas L.J (1998) : "A metric for distributions with applications to image databases", *Computer Vision, Sixth International Conference on 4-jan-1998*, pages 59-66
6. Christos Faloutsos, King-Ip Lin : "Fast Map: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia databases". *SIGMOD'95, Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, pages 163-174

7. Richard Dubes, Anil K.Jain (1979) : “Validity studies in clustering methodologies”. *Pattern Recognition*, Volume 11, Issue 4, pages 235-254
8. Roger N Shepard (1980) : “Multidimensional scaling, tree-fitting and clustering”. *Science, New Series*, Volume 210, Issue 4468, pages 390-398

